Introduction to Computer Vision

Erik G. Learned-Miller
Department of Computer Science
University of Massachusetts, Amherst
Amherst, MA 01003

September 4, 2012

Abstract

NOTE: THIS IS A DRAFT DOCUMENT

Computer vision is the science of endowing computers or other machines with vision, or the ability to see. But what exactly does it mean to see? Most computer vision scientists would agree that seeing is more than the process of recording light in a form that can be played back, like the recording of a video camera. But what, exactly, is needed in addition to the detection or recording of light in order to say that a device, be it natural or manufactured, is seeing?

Perhaps we wish to say that vision is the interpretation of images that leads to actions or decisions, as in the navigation of an autonomous robot. But would we then exclude as vision the process of gazing at the night sky or a beautiful ocean vista, processes in which we may have no intention of making any decision? Processes such as recognition, interpretation, learning, or just enjoyment may be occurring when we see that have no immediate bearing on a decision. On the other hand, something we see may affect a decision we make years later. How do we then know if we are currently seeing or not?

Since vision is a core component of intelligence,[1] its definition encounters many of the same philosophical issues raised when trying to define intelligence itself. Like intelligence, there are many components to vision, including but not limited to memory, retrieval, reasoning, estimation, recognition, and coordination with other senses. It would be odd to insist that all of the above elements be present before we would consider a system to have some degree of vision. At the same time, a system with only one of these abilities might not be promoted to the rank of having vision. To some extent, we define vision by the familiar processes of our own visual systems, and thus, there may be some subjective judgement about the degree to which a system can see by comparing it to our own capabilities.

We will leave the definition in the first sentence of this chapter as it is, and strive to endow computer vision systems with as many capabilities as we can rather than dwelling on whether we have built a system that can truly see.

1 The Purpose of Vision

The purpose of vision in a biological creature is to make inferences about the world from the light impinging upon the creature. Vision is used to find food, to discriminate between a poisonous plant and an edible one, to detect prey and avoid predators, to find shadows to hide in, or to select a mate. Sometimes the inferences lead to immediate action, as in dodging a rock. Other times, we may store the appearance of a scene or object and only act on the visual information at a later time. For example, a squirrel may remember, after finding a nut, that it saw a good place to store it earlier in the day.

Any inference we make about the world through vision may be incorrect, but as humans we are so used to being correct in our visual inferences, it is often hard to believe it when we are wrong. *Visual illusions* are drawings, photographs,

[1] Vision is not *necessary* for intelligence, but is certainly a large part of what a seeing person's brain does.

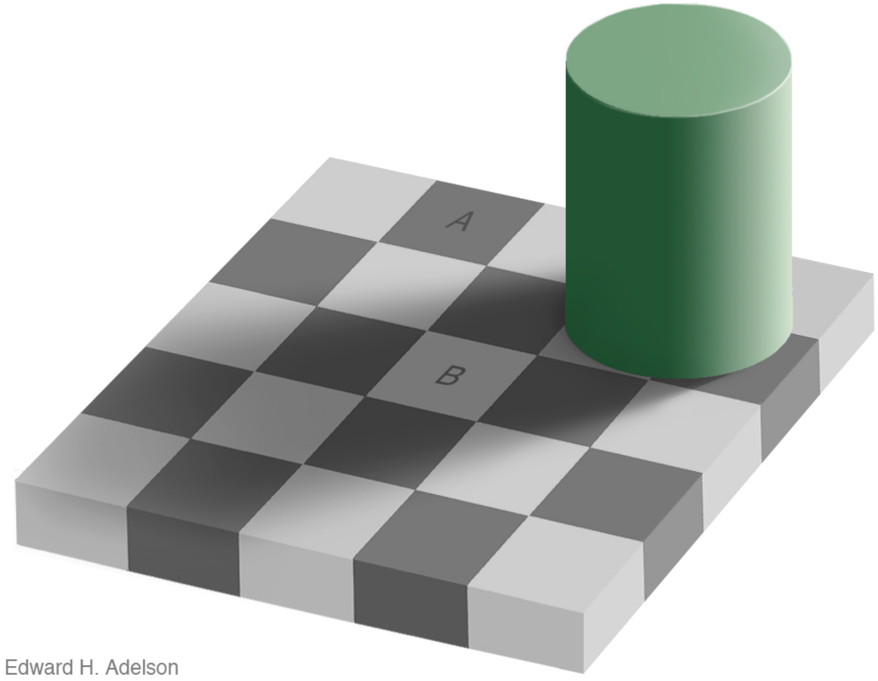

Figure 1: **Checker Shadow Illusion.** The squares marked A and B are the same shade of gray. The illusion is so strong that many people refuse to believe it.

renderings, or other images that highlight the ambiguity and uncertainty associated with the process of visual inference. They are a key tool in understanding visual ambiguity and the processes that humans use to interpret images. By seeing where our interpretations go wrong, we can understand something about how vision happens in people.

Figure 1 shows one of the most well-known visual illusions, created by Edward H. Adelson, a professor of vision science. It illustrates several important phenomena that are central to the understanding of both human vision and computer vision.

Consider the squares in the figure marked A and B. Believe it or not, these squares are rendered with pixels that have the *exact same gray color*. This illusion is so strong that many people simply refuse to believe it. If you are skeptical, see Figure 2 on the next page, which gives your brain an aid in seeing the true brightness of the underlying pixels.

How could a highly sophisticated vision system, namely our own, make such a simple and blundering error? The answer probably lies in the fact that *the human visual system did not evolve to judge differences in the absolute brightness of scene patches.* Rather, it evolved with a higher priority for *judging the*

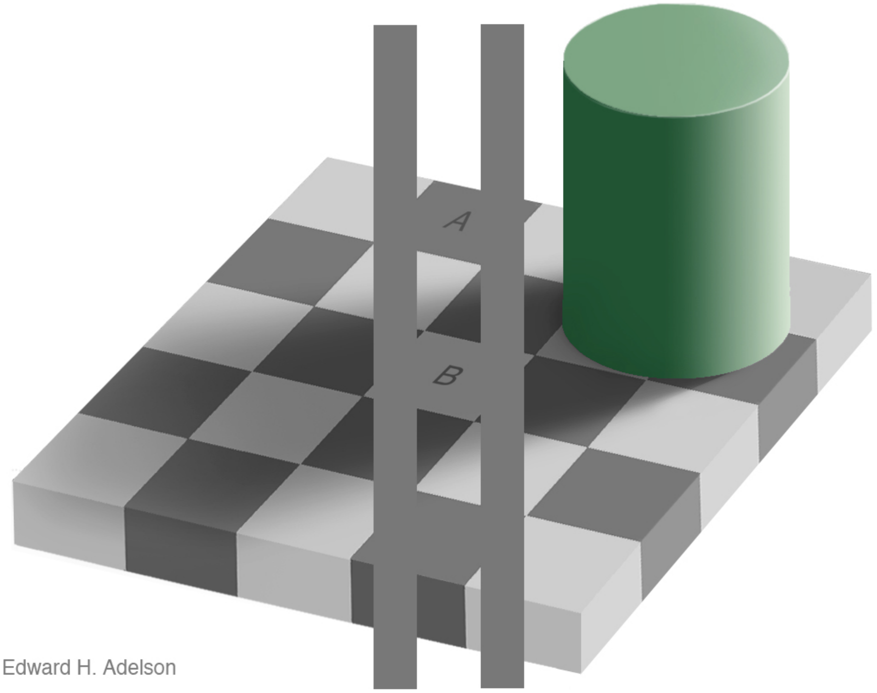

Figure 2: **Checker Shadow Illusion.** The uniformly colored gray bars in front of the scene help to block the effects of the shadow on the brain's interpretation of the scene, allowing us to see that square A and square B are rendered using the same pixel color.

properties of surfaces, after accounting for phenomena such as shadows. The following example illustrates why this might be true.

Consider an herbivore whose diet consists primarily of a particular species of green leafy plant. Suppose that there is a second plant species that is only slightly darker in color than the edible plant, but is highly poisonous. The animal's survival will be closely linked to its ability to discriminate among these plants. While under the same lighting conditions the poisonous plant will reflect less light than the edible plant, we would expect the poisonous plant to reflect *more* light if it is in direct sunlight and the edible plant is in the shade. Thus, the *absolute amount of light* reflected by a surface is a function both of the surface properties *and* of the lighting conditions, and is clearly not a reliable measure of whether the plant is edible. To the animal that is about to eat one of the plants, the important features are the *surface properties*, since they determine the type of plant, and hence whether it is edible or not.

Thus, in situations like this, it is of little relevance how much light a surface is reflecting, but rather, we wish to know about the relative surface properties.

Put another way, we wish to measure a feature of the world that is **invariant** to the specific lighting conditions. The attempt of the visual system to assess properties of the checkerboard that are independent of cast shadows and other lighting phenomena are at least one plausible explanation of the checkered shadow illusion.

The attempt to measure properties of the world that are invariant to various phenomena in which we are not directly interested, such as lighting, is a major theme in computer vision, and we shall encounter it many times. These distracting pheneomena are ofter referred to as *nuisance variables*, and finding ways to interpret the world despite their presence can be a major challenge.

Returning to the central theme of this section, we note that people often *believe* that they are assessing one property of a scene when they are really assessing a very different property. In the checker shadow illusion, a lay person believes they are assessing the amount of reflected light from the scene, when, in reality, they are performing a complicated inference procedure which tries to guess the relative reflectivity of each patch of the scene. From this point of view, we can say that humans are "correct", since, in a realistic embodiment of the scene from Figure 1, they have determined the paint on square A is likely to be darker (i.e., it reflects a lower percentage of the incident light) than the paint on square B. So, considering the *goal of the vision system*, humans are doing something quite reasonable. Computer vision can be as much about figuring out what the answer should be about, e.g. "surface reflectivity", as it can be about figuring out how to get that answer.

Insights into what the human visual system is doing come from many other areas of science including psychology, neuroscience, and ethology (the study of animal behavior). These areas can help us answer questions about what "answers" are useful, and clues about how they might or might not be obtained in animals. In addition to these areas, there are many other fields that are highly relevant to the study of computer vision. We briefly touch on some of these areas in the next section.

2 Related Areas

Computer vision, or from here forward, just vision, is a broad and complex field of study that touches upon many classical fields, and many new areas of inquiry. There are many opinions about what sort of background is necessary for computer vision, but one thing is certain–inspirations for new computer vision methods have come from fields as diverse as psychology, neuroscience, physics, robotics, and statistics. To get a sense of where computer vision lies in relation to some other areas, we briefly describe their overlaps below.

2.1 Optics, Photography, and Photogrammetry

Vision deals with light and its interaction with surfaces, so of course optics plays a role in understanding computer vision systems. Cameras, lenses, focusing, binocular vision, depth-of-field, sensor sensitivity, time of exposure, and other concepts from optics and photography are all relevant to computer vision. Traditionally, when computer vision focused heavily on precise measurements of the world through camera systems, understanding optics was of paramount importance. For better or for worse, as this book is written, the focus on precise measurement using computer vision systems has subsided somewhat, and today the field is more focused on working with uncalibrated systems and noisy measurements.

2.2 Computer Graphics and Art

Computer graphics and art are about making images, whether realistic or fantastic, from knowledge of the world. For example, given a geometric description of a pair of dice, computer graphics algorithms *render* an image of the dice.

Often referred to as the "inverse" of computer graphics, computer vision attempts to make inferences about the world from images. Given a picture of two objects, we would like to infer that they are roughly cubical, and that they are likely to be dice, although we can never be completely sure.

Computer graphics and notions from art can teach us a lot that is useful in computer vision, by making it clear just what cues we use to make inferences about the world. For example, any good portrait artist knows that if a human eye is painted without a "highlight" showing a reflected light in the eye, the person's face can appear lifeless and inanimate. Conversely, a vision system may pick up on subtle specular highlights to conclude that a surface is wet, transparent, or reflective, features associated with living creatures, rather than inanimate objects. By understanding the importance of such cues in making art life-like, we gain insight into the cues that vision systems might use to categorize objects.

2.3 Neuroscience and Physiology

The human eye, the central nervous system, and the brain are all marvels of complex structure and bewildering performance. Studying these systems often provides insight, inspiration, and clues about artificial vision system design. How can a vision system be designed with no external calibration, with no direct measurement of "camera" direction, with no up-front specification of features? The human visual system seems to do all of these things. Even if we are born with sophisticated vision capabilities (which is a source of current debate), we can still ask how the relatively "dumb" process of evolution managed to produce such an extraordinary vision system. Like other evolved capabilities such as flight, we expect to see in simpler organisms precursors of our most sophisticated capabilities that use similar designs, so that evolution or learning could make a small step to produce our current system. These arguments invite a type of analysis that may ultimately lead to more sophisticated artificial vision systems.

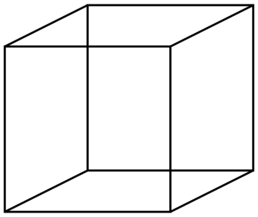

Figure 3: **The Necker Cube.** This classic illustration demonstrates that the same image can result from different real-world objects. In particular, the image shown could result from a wire-frame object in which the viewer's eyes are above the object or a different object in which the viewer's eyes are below the object. If you are having trouble seeing both interpretations of the object, the act of blinking ones eyes often helps to see the other interpretation. Figure 4 gives additional cues to see the two possible orientations of the cube.

2.4 Psychology and Psychophysics

Understanding the limits and capabilities of humans in performing visual tasks can offer important insights into the design of artificial vision systems. Where human vision systems fail dramatically, for example in the presence of certain visual illusions, is a particularly fascinating subject. Human responses to visual illusions can provide insight into processing (such as center-surround filters), deficits (such as the "blind spot" on the retina), and the difference between high-level and low-level visual processing (Kanisza triangle). Psychophysics, a sub-field of psychology that studies how stimuli are perceived by humans and animals, can also offer insights into the structure of processing and assumptions that may be made by humans and other animals. For example, just recording the speed at which a human responds in a particular task, like reading a word, may rule out certain theories as to how certain visual stimuli are processed.

2.5 Probability, Statistics, and Machine Learning

The mathematical subfield of probability, the field of statistics, and the computer science discipline of machine learning have become essential tools in computer vision. Each of these areas plays a major role in computer vision. Here, we make a few introductory comments about the role of these areas of study in computer vision. We will revisit them often as we consider various topics in computer vision.

2.5.1 The Ill-Posed Nature of Vision

We can never be completely sure of what we are seeing, although it certainly doesn't feel that way. The task of vision can be seen as trying to infer the state of the world, or the future state of the world, from the images that fall upon our retinas. Since there are many different states of the world that could produce the same images on our retinas, there is no fool-proof way of distinguishing among the various structures or objects that might have created a particular image. Thus, if our goal is to infer the state of the world with certainty, then we are defeated from the start. In mathematical terms, vision is an *ill-posed* problem, since there does not exist a single, correct answer to the question of questions like "What kind of object is pictured in this image?" This, like many other phenomena in computer vision, is highlighted by certain classical visual illusions like the Necker cube, shown in Figure 3.

Because there are multiple potential causes of each image we see, it is helpful to be able to select some notion of the "best" one. While there are many potential methods for deciding which one is best, a common approach is to try the following.

1. Develop a simplified statistical model of the experimental setting.

2. Using the statistical model, evaluate the probability of each outcome.

3. Choose the outcome that is consistent with our observations whose probability is highest under the statistical model.

Following these steps is a complex process that embodies much of the work done by computer vision researchers today. There is no "best" statistical model for a particular problem. Different models make different assumptions in an attempt to either run faster, give more accurate answers, be applicable in more general settings, or satisfy various other requirements.

To illustrate the idea, a model might assert the probabilty of seeing an object from above is higher than the probability of seeing something from below (since humans have a slightly downward gaze most of the time). Since a major difference between the two views of a Necker cube can be explained by whether one is looking up at a cube or down at a cube, under this type of model, the version of the left side of Figure 4 would be the interpretation with the highest probability.

In this book we will develop a number of statistical models, and as we do so we will try to comment on their relative strengths and weaknesses from a variety of different viewpoints.

2.5.2 Limitations of probability and statistics

While techniques of probability and statistics may appear to be useful in computer vision, they are certainly not a panacea. A variety of troubling questions remain for which there are not yet any good answers. Some of these questions include the following.

The stationarity assumption. Many machine learning methods assume that the distributions from which we are "trained" are the same distributions on which we are "tested". In other words, they assume that the probability of something occurring in the past is the same as that of it occurring in the future.

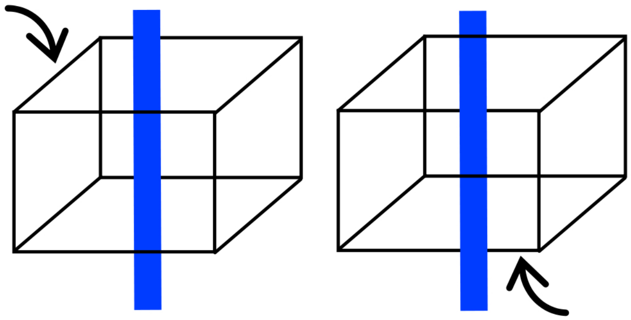

Figure 4: **Disambiguating the Necker Cube.** One cue that can disambiguate the geometry behind an ambiguous image is *occlusion*. On the left, the occlusions imply a certain orientation of the Necker cube in which the observer is looking down on the top of a cube. On the right, the occlusions suggest that we are looking up into the bottom of a cube.

We will refer to this as the *stationarity assumption*, since stationarity is the property of a random process, informally speaking, that its parameters do not vary in time. However, it is rare to encounter a true process that is stationary, or even approximately stationary. Thus, some of the fundamental assumptions that are behind many learning algorithms appear to be false.

Insufficient training data. If we are to base our decisions on visual experience, it would appear that we should estimate the probabilities of complex events that we see in images. But the number of degrees of freedom in complex stimuli, such as faces, is enormous, and the number of examples needed to estimate distributions of quantities with large numbers of degrees of freedom is enormous. It frequently appears as though *there is not enough data* to estimate the probabilities of interest well. If these probabilities are estimated poorly, then we expect the decisions based on them to be poor as well. What is the resolution of this paradox?

Developing models with limited degrees of freedom appears to be one possible route out of this quagmire. Techniques such as regularization, sparseness priors, manifold learning, and feature selection, some of which are addressed in this text, are all attempts to deal with this problem. But up to now, they are mostly quite unsatisfying. We still do not understand how humans are capable of learning from data as efficiently as they do.

3 Goals of Computer Vision

In addition to the many fields with which it intersects, computer vision is complicated by the highly varying agendas of its practitioners. People working on the same problem, such as face recognition, may have very different approaches, not because they disagree on fundamental principles, but more because they have different goals. In this section, we examine computer vision from a variety of different perspectives as defined by goals of different researchers.

3.1 Computer Vision as Engineering

A good deal of the field is focused on developing applications that can be used in the real world. Some examples include quality control in manufacturing, optical character recognition, driver assistance systems, surveillance, photography and entertainment.

At the risk of oversimplifying the discussion, we will refer to this as the "engineering" approach to computer vision. The goal in this approach is to make things work in the near term. This work in general is characterized by

- solving real-world problems in need of a solution, rather than "toy" problems invented by the researcher;
- making vision methods fast enough to be useful, or faster so that they are more useful;
- making vision systems more robust, so that they work in a wider range of environments; and
- designing systems using currently available technology, so that it is easier to predict the successful completion of specific projects.

Sometimes the engineering approach carefully specifies a narrow application domain, and builds a highly specialized application which would fail in any other domain, but works very well in the specified domain. For example, techniques for analyzing printed circuit boards for flaws rely on careful alignment of the target board with respect to a video camera, a fixed lighting arrangment, and assumptions about the type of camera used to collect the data. Such product inspection systems represent one highly successful area of computer vision deployed algorithms. However, few people would expect such algorithms to be useful to an autonomous robot for recognizing faces. They simply weren't designed for the same thing.

There are many practical vision problems that we are not yet able to solve. For example, there would be many applications of a program that could look at a photograph and name the people in it. While we have made some progress on this problem, we are not nearly as good at it as people are yet.

Given that there are certain problems for which we cannot yet engineer a sufficiently high quality solution, the question emerges about how to best proceed towards a solution. One method is to focus on engineering systems that are as good as possible, and try to make incremental progress on an easily quantifiable measure of fitness, such as accuracy on a face recognition task.

Another approach is to study the central principles of computer vision and natural vision systems and then build fundamentally new systems using this new understanding. The danger, of course, is that we never return from these fundamental investigations of core principles to build a useful vision system.

Nevertheless, there are a large number of vision researchers who are focused on understanding the principles behind vision rather than producing short-term artifacts.

3.2 Computer Vision as a Route to Understanding Intelligence

The human brain is a good candidate for the most complex and intriguing structure ever encountered. It is one of the most familiar and least understood structures known. Even the top neurologists, neuroscientists, psychologists, philosophers, and computer scientists are baffled at the capabilities of the human brain and human vision.

There are many ways to study the brain, to study human intelligence, to study behavior. Philosophers study fundamental questions such as whether it is possible for a machine to be conscious, or for a human to have free will in a deterministic universe. Psychologists form general theories of behavior and assess them in new scenarios to test their predictive power. Neuroscientists dissect cadaver brains and implant electrodes into living creatures, including monkeys and humans having brain surgery, to record the activity of single neurons. All of these fields have offered invaluable insights into human behavior and the workings of the brain.

Additional insights can be gained by trying to *build* something that works like, or works as well as, or perhaps even works better than the brain in solving certain problems. Developing real vision systems

- gives us insight into which problems are easy and which problems are hard;
- allows us to investigate the limits of "low level" learning, high level learning, and context;
- forces us to deal with difficult practical issues of representation;
- makes us consider the vast memory of the brain, its limitations, and its compromises.

Working on vision systems may not provide precise answers to any of these problems, but it certainly forces one to deal with these issues. And the results are often surprising, and vastly different from what might conclude using other methods of inquiry. This means that computer vision is a great complement to other ways of understanding intelligence. In the end, it seems likely that a thorough and deep understanding of the brain will require significant contributions from all of the areas mentioned above.